Claims

What is claimed is:

1. Apparatus for presenting images representative of one or more words in an utterance with corresponding decoded speech, the apparatus comprising:

a visual detector, the visual detector capturing images of body movements corresponding to one or more words in the utterance;

a visual feature extractor coupled to the visual detector, the visual feature extractor receiving time information from an automatic speech recognition (ASR) system and operatively processing the captured images into one or more image segments based on the time information relating to one or more words, decoded by the ASR system, in the utterance, each image segment comprising a plurality of successive images in time corresponding to a decoded word in the utterance; and

an image player operatively coupled to the visual feature extractor, the image player receiving and presenting each image segment with the corresponding decoded word.

2. The apparatus of claim 1, wherein the image player repeatedly presents one or image segments with the corresponding decoded word.

3. The apparatus of claim 1, further comprising:

a delay controller operatively coupled to the visual feature extractor, the delay controller selectively controlling a delay between an image segment and a corresponding decoded word in response to a control signal.

4. The apparatus of claim 1, further comprising:

a visual detector for monitoring a position of a user;

a position detector coupled to the visual detector, the position detector comparing the position of the user with a reference position and generating a control signal, the control signal being a first value when the position of the user is within the reference area and being a second value when the position of the user is not within the reference area;

a label generator coupled to the position detector, the label generator displaying a visual indication on a display in response to the control signal from the position detector.

5. The apparatus of claim 4, wherein the label generator receives information from the ASR system, the label generator using the information from the ASR system to operatively position the visual indication on the display.

6. The apparatus of claim 1, wherein the body movements include at least one of lip movements of the speaker, mouth movements of the speaker, hand movements of a sign interpreter of the speaker, and arm movements of the sign interpreter of the speaker.

7. The apparatus of claim 1, further comprising:

a display controller, the display controller selectively controlling one or more characteristics of a manner in which the image segments are displayed with corresponding decoded speech text.

8. The apparatus of claim 7, wherein the display controller operatively controls at least one of a number of times an image segment animation is repeated, a speed of image animation, a size of an image segment on a display, a position of an image segment on the display, and a start time to process a next image segment.

9. The apparatus of claim 1, wherein the image player displays each image segment in a separate window on a display in close proximity to the decoded speech text corresponding to the image segment.

10. Apparatus for presenting images representative of one or more words in an utterance with corresponding decoded speech, the apparatus comprising:

an automatic speech recognition (ASR) engine for converting the utterance into one or more decoded words, the ASR engine generating time information associated with each of the decoded words;

a visual detector, the visual detector capturing images of body movements corresponding to one or more words in the utterance;

a visual feature extractor coupled to the visual detector, the visual feature extractor receiving the time information from the ASR engine and operatively processing the captured images into one or more image segments based on the time information relating to the decoded words, each image segment comprising a plurality of successive images in time corresponding to a decoded word in the utterance; and

an image player operatively coupled to the visual feature extractor, the image player receiving and presenting each image segment with the corresponding decoded word.

11. The apparatus of claim 10, wherein the image player repeatedly presents one or image segments with the corresponding decoded word.

12. The apparatus of claim 10, further comprising:

a delay controller operatively coupled to the visual feature extractor, the delay controller selectively controlling a delay between an image segment and a corresponding decoded word in response to a control signal.

13. A method for presenting images representative of one or more words in an utterance with corresponding decoded speech, the method comprising the steps of:

capturing a plurality of images representing body movements corresponding to the one or more words in the utterance;

associating each of the captured images with time information relating to an occurrence of the image;

receiving, from an automatic speech recognition (ASR) system, data including a start time and an end time of a word decoded by the ASR system;

aligning the plurality of images into one or more image segments according to the start and stop times received from the ASR system, wherein each image segment corresponds to a decoded word in the utterance; and

presenting an image segment with a corresponding decoded word.

14. The method of claim 13, further comprising the step of:

selectively controlling a delay between when an image segment is presented and when a decoded word corresponding to the image segment is presented.

15. The method of claim 13, further comprising the step of:

selectively controlling a manner in which an image segment is presented with a corresponding decoded word.

16. The method of claim 13, further comprising the steps of:

monitoring a position of a user;

comparing the position of the user with a reference position and generating a control signal having a first value when the position of the user is within the reference position and a second value when the position of the user is outside of the reference position; and

presenting a visual indication on a display screen in response to the control signal.

17. The method of claim 13, wherein the step of aligning the plurality of images comprises:

comparing the time information relating to the captured images with the start and stop times for a decoded word; and

determining which of the plurality of images occur within a time interval defined by the start and stop times of the decoded word.

18. In an automatic speech recognition (ASR) system for converting an utterance of a speaker into one or more decoded words, a method for enhancing the ASR system comprising the steps of:

capturing a plurality of successive images in time representing body movements corresponding to one or more words in the utterance;

associating each of the captured images with time information relating to an occurrence of the image;

obtaining, from the ASR system, time ends for each decoded word in the utterance;

grouping the plurality of images into one or more image segments based on the time ends, wherein each image segment corresponds to a decoded word in the utterance; and

presenting an image segment with a corresponding decoded word.

19. The method of claim 18, wherein the step of obtaining time ends for a decoded word from the ASR system comprises determining a start time and a stop time associated with the decoded word.

20. The method of claim 18, wherein the step of grouping the plurality of images into image segments comprises:

comparing the time information relating to the captured images with the time ends for a decoded word; and

determining which of the plurality of images occur within a time interval defined by the time ends of the decoded word.

21. The method of claim 18, wherein the step of presenting an image segment with a corresponding decoded word comprises repeatedly displaying the image segment as an animation of successive images in time.

22. The method of claim 21, further comprising the step of presenting the image segment in a separate window on a display screen with the decoded word corresponding to the image segment.

23. The method of claim 18, wherein the body movements captured in the images include at least one of lip movements of the speaker, mouth movements of the speaker, hand movements of a sign interpreter of the speaker, and arm movements of the sign interpreter of the speaker.

24. A method for presenting images representative of one or more words in an utterance with corresponding decoded speech, the method comprising the steps of:

providing an automatic speech recognition (ASR) engine;

decoding, in the ASR engine, the utterance into one or more words, each of the decoded words having a start time and a stop time associated therewith;

capturing a plurality of images representing body movements corresponding to the one or more words in the utterance;

buffering the plurality of images by a predetermined delay;

receiving, from the ASR engine, data including the start time and the end time of a decoded word;

aligning the plurality of images into one or more image segments according to the start and stop times received from the ASR engine, wherein each image segment corresponds to a decoded word in the utterance; and

presenting an image segment with a corresponding decoded word.